

Evolution of Tokenization: Handling Out-of-Vocabulary Words

Input Text:

“The antidisestablishmentarianism was extraordinary”

1. Traditional Word-Level

(vocab_size = 5,000)



50% UNK rate

Information loss

3. Subword Tokenization

(WordPiece/BPE)



0% UNK rate

Adaptive handling

2. Character-Level

(no vocabulary limit)



0% UNK rate

Very long sequences

4. Hybrid Approach

(Subword + Character fallback)

Comparison Metrics				
Method	UNK Rate	Seq Length	Efficiency	Coverage
Traditional	High	Short	Fast	Poor
Character	None	Very Long	Slow	Perfect
Subword	Minimal	Medium	Good	Excellent
Hybrid	None	Variable	Adaptive	Perfect

Evolution toward better OOV handling

Key Insight:
Subword tokenization
balances all factors

Modern Trend:
Hybrid approaches for
ultimate flexibility